WELCOME TO
R - ROGRAMMING LANGUAGE

Regression Using R Language

- Language:

ntroduction:

* *R is a programming and free software nvironment for statistical computing and graphics.*

* *An effective data handling and storage facility.*

* *A large,coherent,integrated collections of termediate,tools for data analysis.*

* *Programming language includes conditions ops,user-defined,recursive functions and input & output cilities.*

egression:

tro:

* *It is introduced by "Sir Francis Galton".*

* *It means "Stepping back towards the average".*

* *Regression analysis the mathematical measure of e average relationship between two or more variables terms of the original units of the data.*

* *Estimation of regression is called regression analysi*

egression in R Language:

Regression analysis is a widely used statistical tool to establish a relation model between two variables.

One of these variable is called "predictor variable" whose value is gathered through experiments.

The other variable is called "response variable" whose value is derived from the predictor variable.

egression are two types:

1. *Linear Regression*
2. *Multiple Regression*

near Regression in R:

In linear regression these two varibles are related throug an equation, where exponent(power) of both these variables is 1.

Mathematically a linear relationship represents a straight line when plotted as a graph.

A non-linear relationship where the exponent of any variable is not equal to 1 creates a curve.

ormula for Linear Regression:

The general mathematical equation for a linear regression is-

$$Y = a + bX$$

$$Y = (\overline{Y} - b\overline{X}) + bX$$

$$Y = \overline{Y} + b(X - \overline{X})$$

$$Y = \overline{Y} + bx$$

ollowing is the description of the parameters used-

* *Y is the response variable.*

* *X is the predictor variable.*

* *a and b are constants which are called the*
oefficients.

* *bxy is the parameter of regression*

The Regression Model

Dependent variable

Y-intercept

Slope coefficient

Independent variable

Random error term

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Systematic component

Random component

teps to Establish a Regression:

rry out the experiment of gathering a sample of observed ues of height and corresponding weight.

eate a relationship model using the lm() functions in R.

d the coefficients from the model created and create the thematical equation using these.

t a summary of the relationship model to know the averag or in prediction.Also called residuals.

predict the weight of new persons, use the predict() funct R.

nput data:

btain the equation of two variables of regression fo
e following data and also find out the estimation o
of x=180.

of	176	154	148	166	172	124	190	135	155
f	88	61	59	70	88	65	92	52	65

) Function:

The function creates the relationship model bew

redictor and the response variable.

ntax:

The basic syntax for lm() function in linear regres

lm(formula,data)

Following is the description of the parameters used:

* *formula* - symbol presentation the relation betwe

.

* *data* - vector on which the formula will be applie

orrelation Coefficient:

The correlation coefficient between two random variable and Y is defined as

$$Pxy = corr\ (X,Y) = cov(X,Y) \backslash \sigma x \sigma y$$

here ,

$$cov(X,Y) = E\ (X-\mu x)(Y-\mu y) \backslash \sigma x \sigma y$$

t has a value -1 and +1,and it indicates the degree of line ependence between the variables. It detects only linear ependence between two variables.

e estimate interpretation when both variables are nous:

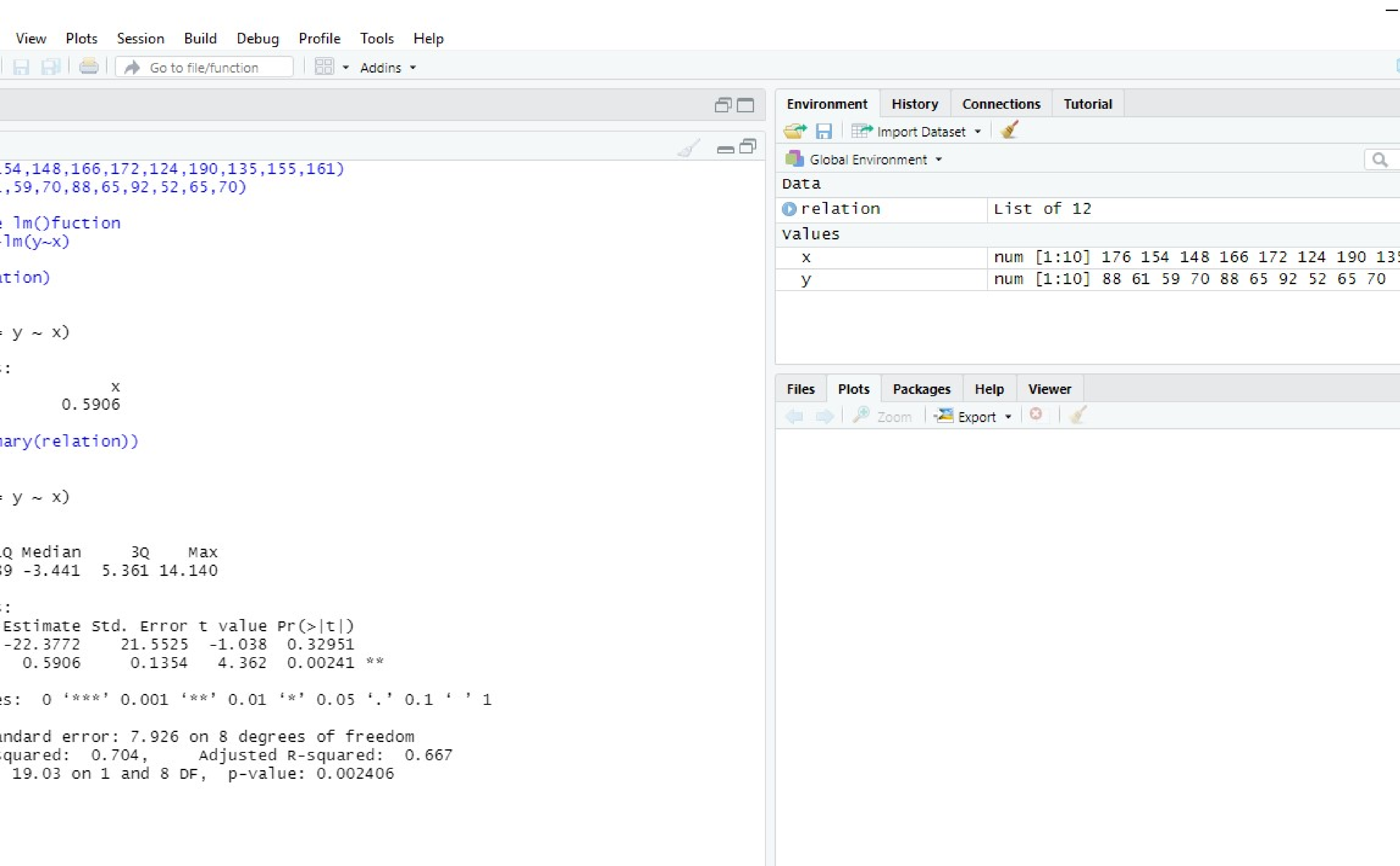
Given a one unit increase in X,this is the expected ge in Y,on average.

(This interpretation changes for categorical variables and variable transformation)

dard Error:

The standard error is the estimated variability in a ficient due to sampling variability i.e.a different sample ma ts in different coefficients and the variability of coefficient s samples is estimated bt the standard error of the ective coefficient.

ut:


View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
.54,148,166,172,124,190,135,155,161)
.,59,70,88,65,92,52,65,70)
e lm()fuction
-lm(y~x)
ation)
= y ~ x)
x
0.5906
mary(relation))
= y ~ x)
1Q Median 3Q Max
39 -3.441 5.361 14.140
Estimate Std. Error t value Pr(>|t|)
-22.3772 21.5525 -1.038 0.32951
0.5906 0.1354 4.362 0.00241 **
es: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
andard error: 7.926 on 8 degrees of freedom
squared: 0.704, Adjusted R-squared: 0.667
19.03 on 1 and 8 DF, p-value: 0.002406
Environment History Connections Tutorial
Import Dataset
Global Environment
Data
relation List of 12
Values
x num [1:10] 176 154 148 166 172 124 190 13
y num [1:10] 88 61 59 70 88 65 92 52 65 70
Files Plots Packages Help Viewer
Zoom
Export

ict() function:

ax:

The basic syntax for predict() in linear regression is-

predict(object,newdata)

Following is the description of the parameters used-

* *object - formula which is already created using li*

on.

* *new data - vector containing the value for predic*

ble.

utput:

redict the weight of a person & given x(height) = 180:

Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

~/

```
 predictor vector
(176,154,148,166,172,124,190,135,155,161)

 response vector
(88,61,59,70,88,65,92,52,65,70)

ly the lm()fuction
tion<-lm(y~x)

d the weight of a person with height 180
ata.frame(x=180)
lt<-predict(relation,a)
t(result)
 1
61
```

isualize the regression Graphically:

Studio

Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

urce

nsole ~/

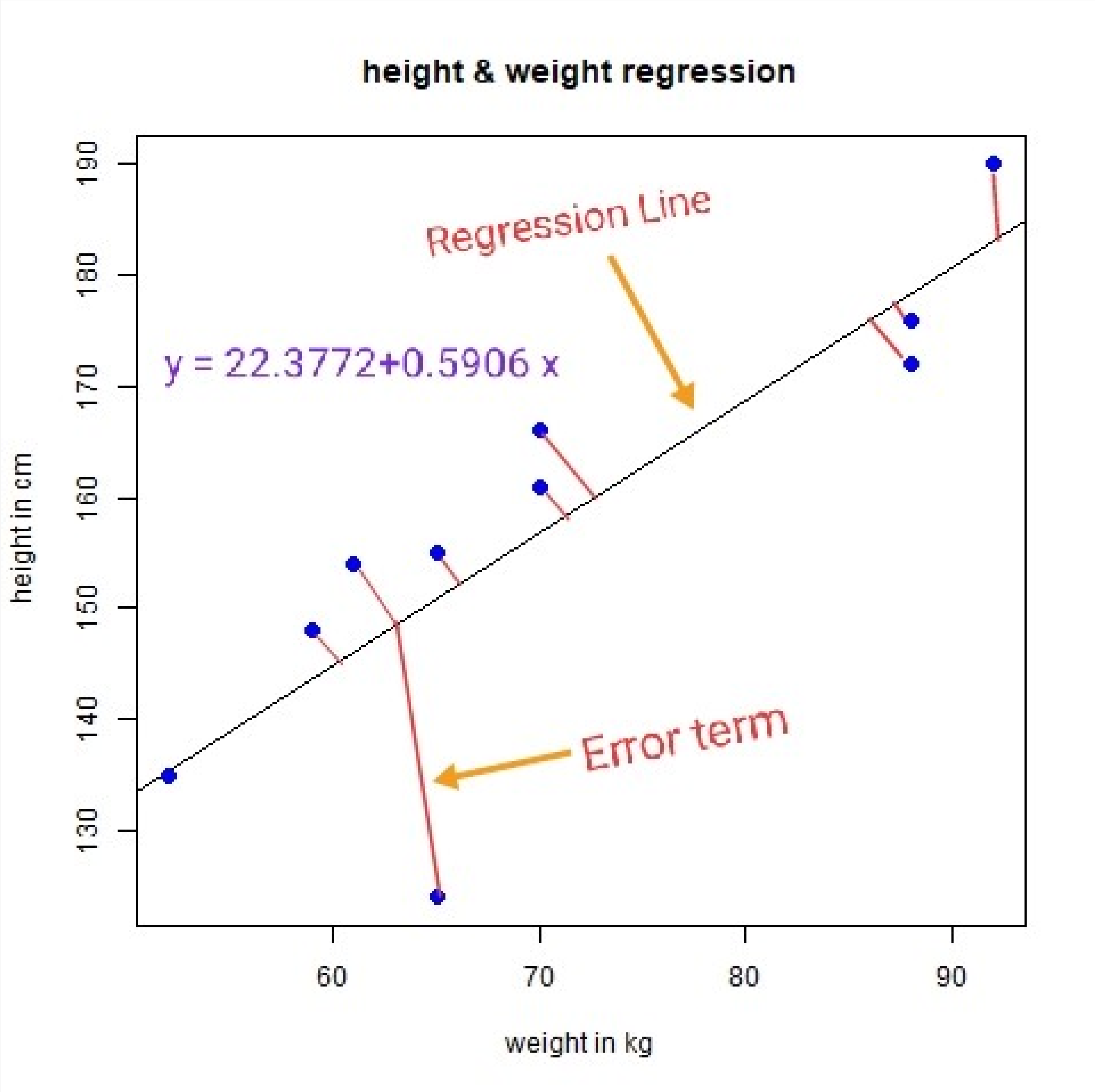
```
#create the predictor and response variable
x<-c(176,154,148,166,172,124,190,135,155,161)
y<-c(88,61,59,70,88,65,92,52,65,70)
relation<-lm(y~x)

#give the chart file a name
png(file="linearregression.png")

#plot the chart.
plot(y,x,col="blue",main="height & weight regression",abline(lm(x~y)),cex=1.3,pch=16,xl
="weight in kg",ylab="height in cm")

#save the file
dev.off()
ll device
        1
```

raph:



height & weight regression
Regression Line
y = 22.3772+0.5906 x
Error term
height in cm
weight in kg
190
180
170
160
150
140
130
60
70
80
90

esult:

Regression equation of y on x;

$$y = -22.3772 + 0.5906\, x$$

The predict value,

height of x =180 ,then weight of y = 83.934

THANK
YOU

MULTIPLE REGRESSION USING R_LANGUAGE

EGRESSION

- Regression analysis is used to establish a relationship model between two variables.
- One of these variable is called independent variable whose value is gathered through experiments.
- The other variable is called dependent variable whose value is derived from the independent variable.
- Formula for regression

$$Y = \alpha + \beta X$$

ULTIPLE REGRESSION

Multiple regression is an extension of linear regression into relationship between more than two variables.

In simple linear relation we have one dependent and one independent variable, but in multiple regression we have more than one independent variable and one dependent variable.

ORMULA FOR MULTIPLE REGRESSION

The general mathematical equation for multiple equation is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta n X n$$

- **Y** is the dependent variable
- α, β_1, $\beta_2 \ldots, \beta_n$ are the parameter
- X_1, X_2, $\ldots, X_n$ are the independent variables

TEPS TO ESTABLISH A MULTIPLE REGRESSION

Step 1: **Collect the data**

Step 2: Capture the data in R

Step 3: Check for linearity

Step 4: Apply the multiple regression in R

Step 5:Make a prediction

e syntax for multiple regression:

$$\text{lm}(y \sim x_1+x_2+x_3...,\text{data})$$

() Function:

This function creates the relationship model between the pendent and the Independent variable.

input

o let's start with a simple example where our goal is to predict the ock_index_price (the dependent variable) of a fictitious economy based two independent/input variables:

YEAR	2020	2020	2020	2020	2020	2020	2020	2020	2020
MONTH	10	9	8	7	6	5	4	3	2
REST RATE	2.75	2.75	2.5	2.5	2.25	2.25	2.25	2	2
MPLOYMENT RATE	5.3	5.3	5.3	5.3	5.4	5.6	5.5	5.5	5.5
OCK INDEX PRICE	1464	1394	1357	1293	1256	1254	1234	1195	1159

Check for linearity

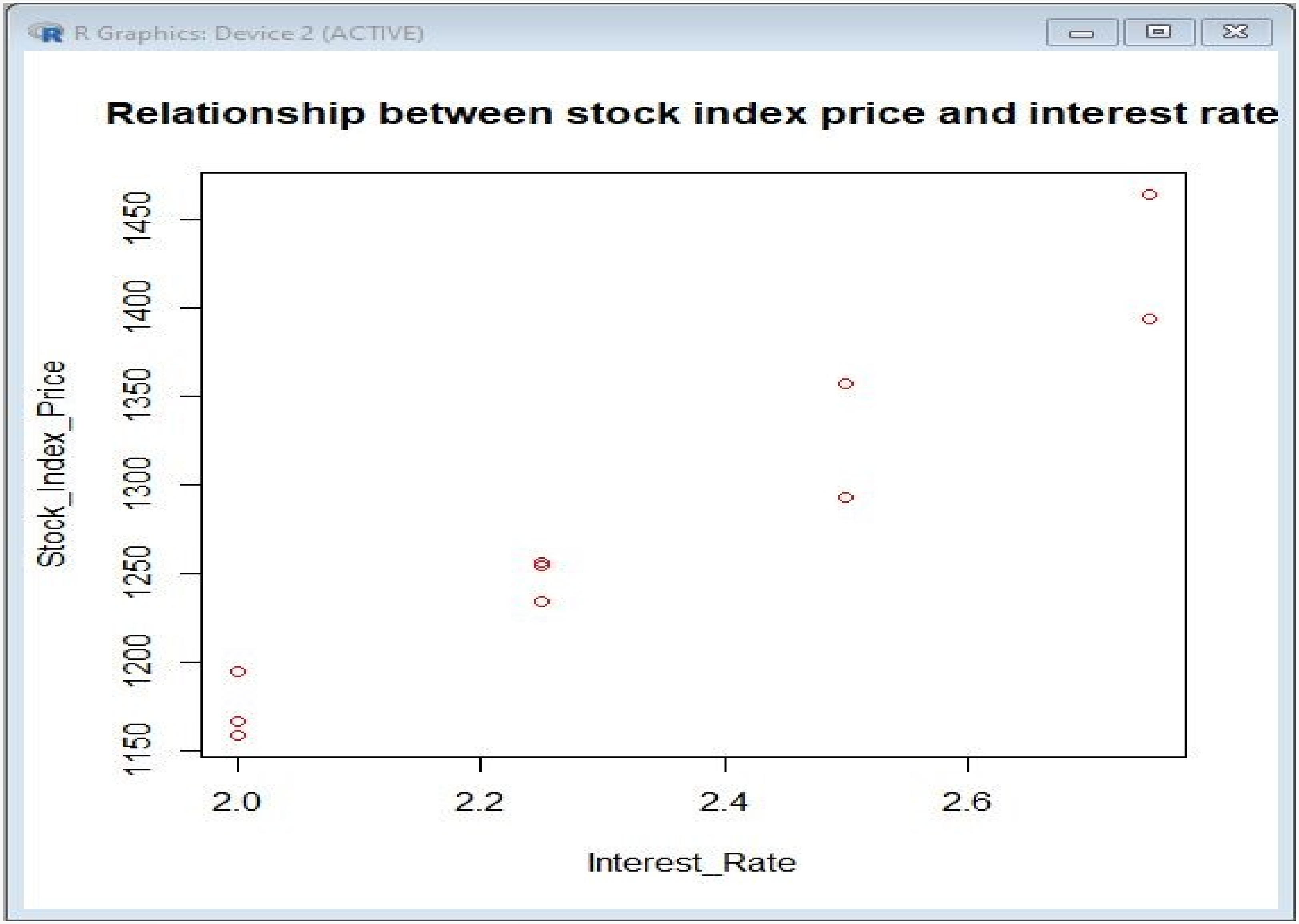
Before you apply linear regression models. Most notably, you'll need to r
hat a linear relationship exists between the dependent variable and the
endent variables.A quick way to check for linearity is by using scatter plots.

ur example, we'll check that a linear relationship exists between:
he Stock_Index_Price (dependent variable) and the Interest_Rate (indepen
riable)
he Stock_Index_Price (dependent variable) and the Unemployment_Rate
ndependent variable)

ntax that can be used in R to plot the relationship between
e Stock_Index_Price and the Interest_Rate:

```
Year <- c(2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020,
2020)
Month <- c(10,9,8,7,6,5,4,3,2,1)
Interest_Rate <- c(2.75,2.75,2.5,2.5,2.25,2.25,2.25,2,2,2)
Unemployment_Rate <- c(5.3,5.3,5.3,5.3,5.4,5.6,5.5,5.5,5.5,5.6)
Stock_Index_Price <-
c(1464,1394,1357,1293,1256,1254,1234,1195,1159,1167)
plot(x=Interest_Rate, y=Stock_Index_Price, main='Relationship
between stock index price and interest rate‘, col="red”)
```

Scatter plot for relation betweem interest rate and stock index price

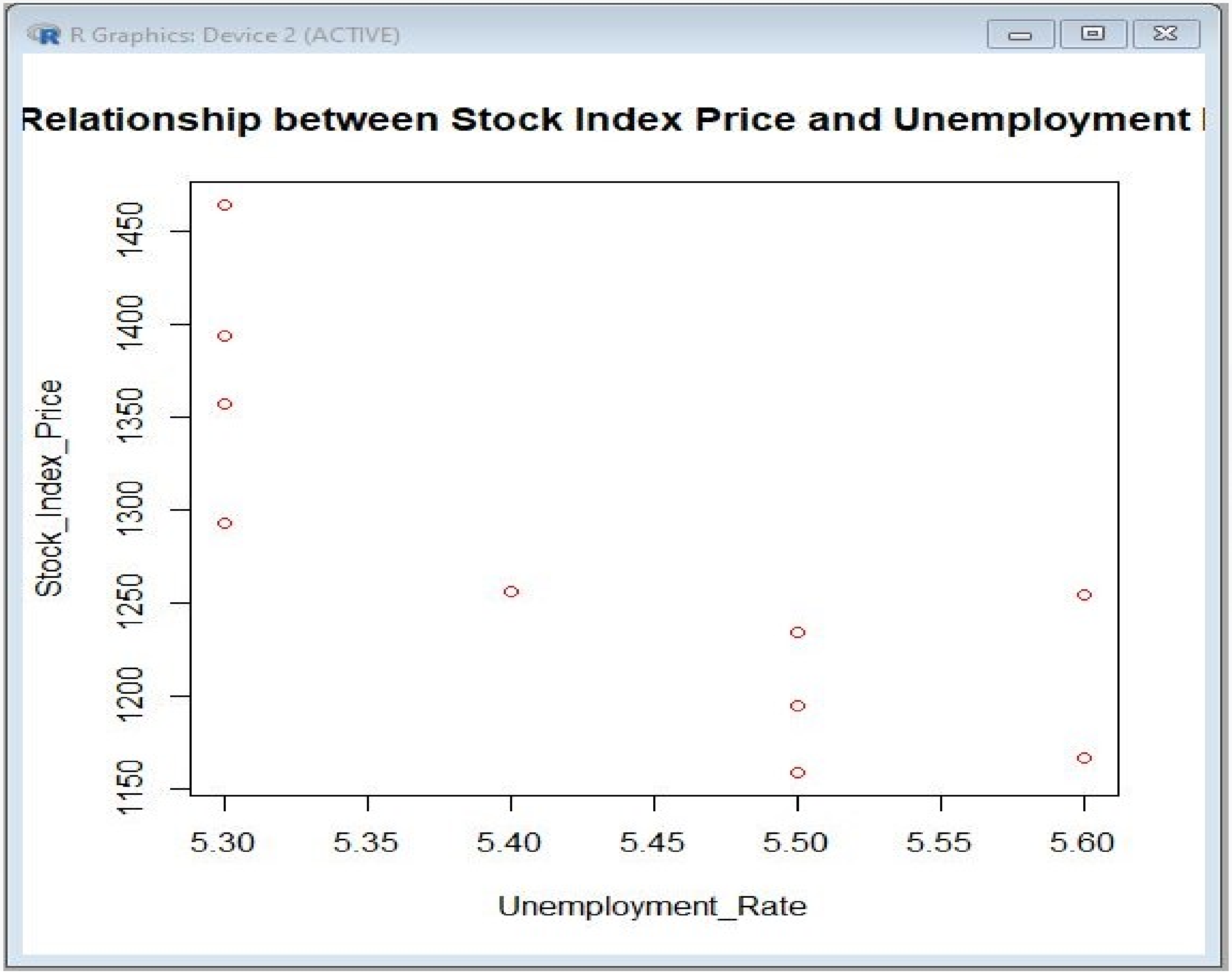


r the second case, you can use the syntax below in order to plot the ationship between the Stock_Index_Price and the Unemployment_Rate:

```
ar <- c(2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020)
nth <- c(10,9,8,7,6,5,4,3,2,1)
erest_Rate <- c(2.75,2.75,2.5,2.5,2.25,2.25,2.25,2,2,2)
employment_Rate <- c(5.3,5.3,5.3,5.3,5.4,5.6,5.5,5.5,5.5,5.6)
ck_Index_Price <- c(1464,1394,1357,1293,1256,1254,1234,1195,1159,1167)
t(x=Unemployment_Rate, y=Stock_Index_Price,main='Relationship between Stock Index
ce and Unemployment Rate',col="red")
```

Scatter plot for relation betweem unemployment rate and stock index price

inear relationship
ts between the
dex_Price and the
oyment_Rate .
e unemployment
up, the stock
ce goes down
still have a linear
hip, but with a
slope)



pply the multiple regression in R

g the syntax for our example:

```
ear <- c(2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020)
onth <- c(10,9,8,7,6,5,4,3,2,1)
terest_Rate <- c(2.75,2.75,2.5,2.5,2.25,2.25,2.25,2,2,2)
nemployment_Rate <- c(5.3,5.3,5.3,5.3,5.4,5.6,5.5,5.5,5.5,5.6)
tock_Index_Price <- c(1464,1394,1357,1293,1256,1254,1234,1195,1159,11
odel <- lm(Stock_Index_Price ~ Interest_Rate + Unemployment_Rate)
ummary(model)
```

output

e you run the code in R Language, you'll get the following out

```
Call:
lm(formula = Stock_Index_Price ~ Interest_Rate + Unemployment_Rate)

Residuals:
    Min      1Q  Median      3Q     Max 
-42.483 -16.130  -0.442  17.183  44.216 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        458.707    894.398   0.513 0.623828    
Interest_Rate      337.205     61.374   5.494 0.000912 ***
Unemployment_Rate    6.371    142.138   0.045 0.965502    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.41 on 7 degrees of freedom
Multiple R-squared:  0.9334,    Adjusted R-squared:  0.9143 
F-statistic: 49.04 on 2 and 7 DF,  p-value: 7.631e-05
```

ummary

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

Stock Index Price = (Intercept) + (Interest Rate coefficent)*X_1 + (Unemployment Rate coefficent)*X_2

Multiple Regression fit for Stock Index Price:

Stock Index Price = (458.707) + (337.205)*X_1 + (6.371)*X_2

ake a prediction

r example, imagine that you want to predict the stock index price ter you collected the following data:

terest Rate = 1.5 (i.e., X_1= 1.5)

nemployment Rate = 5.8 (i.e., X_2= 5.8)

nd if you plug that data into the regression equation you'll get:

- **Stock Index Price = (458.707) + (337.205)*1.5 + (6.371)*5.8**

ne predicted value for the Stock Index Price is therefore 1001.4663